

开源长期保存系统DAITSS研究

吴振新 向菁
国家科学图书馆

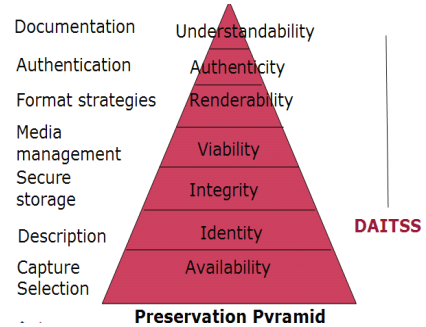
数字资源长期保存系统

项目来源 ^①	Fedora ^②	DSpace ^③	LOCKSS ^④	DAITSS ^⑤	E-prints ^⑥	DIAS ^⑦	FAMDS ^⑧
开发 ^①	Fedora ^②	DSpace ^③	LOCKSS ^④	FDA项目 ^⑤	E-prints ^⑥	e-depot ^⑦	FAMORA ^⑧
开放源代码许可协议 ^①	ECL 1.0 ^②	BSD ^③	BSD ^④	GNU GPL ^⑤	GNU GPL ^⑥	^⑦	^⑧
最新版本发布日期 ^①	2008.1 ^②	2008.9 ^③	^④	2008.8.29 ^⑤	2008.11 ^⑥	^⑦	2007.1 ^⑧
最新版本号 ^①	3.1 ^②	1.5.1 ^③	1.96.8 ^④	2.0 ^⑤	3.1.2-1 ^⑥	2.0 ^⑦	3.0 ^⑧
适用操作系统 ^①	UNIX/MacOSX/Windows ^②	UNIX//MacOSX//Windows ^③	Linux/Solaris ^④	Unix//Linux ^⑤	UNIX/Linux/Solaris ^⑥	^⑦	^⑧
是否有社区 ^①	Fedora commons社区 ^②	Wiki ^③	邮件列表 ^④	Wiki ^⑤	Wiki ^⑥	从属项目网站 ^⑦	从属项目网站 ^⑧

DAITSS简介

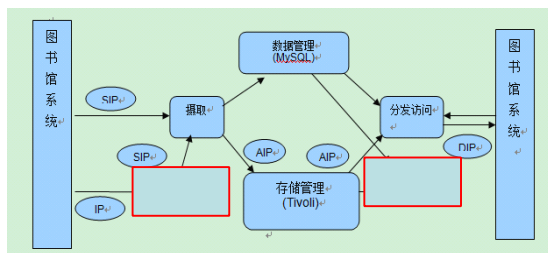
- 佛罗里达图书馆自动化中心（FCLA）为佛罗里达数字保存系统（FDA）开发
- 优势：
 - 专注于保存功能：Dark Archive In The Sunshine State（黑暗存档系统）
 - 遵循当前已有的开放标准和技术规范
 - ✓ 遵循OAIS
 - ✓ METS
 - ✓ 全面兼容PREMIS
 - ✓ Z39.87（数字静态图像技术元数据规范）
 - 积极长期保存策略：支持格式规范、大规模格式迁移和按需迁移等

DAITSS专注于保存功能，不支持外部用户的直接访问。



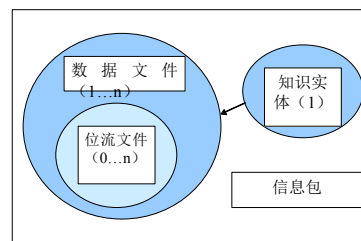
Priscilla Caplan, FCLA Digital Archive, 2005 Joint Meeting of the CSUL Committees.

系统功能框架



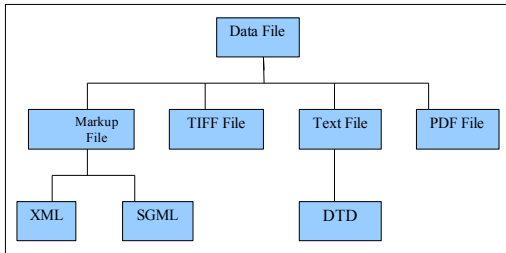
Reference: DAITSS Operations Manual. <http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf>

数字对象模型



- ✓ 知识实体 (Intellectual Entity)
- ✓ 数据文件 (Data File)
- ✓ 比特流 (Bit stream)

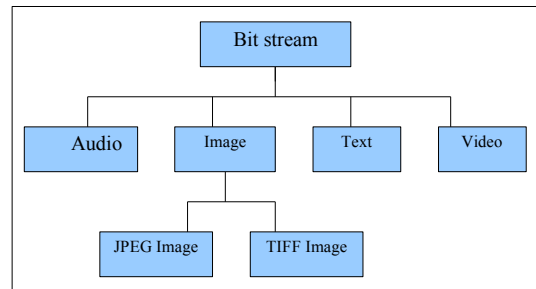
DAITSS 数据文件对象



第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

DAITSS 比特流对象

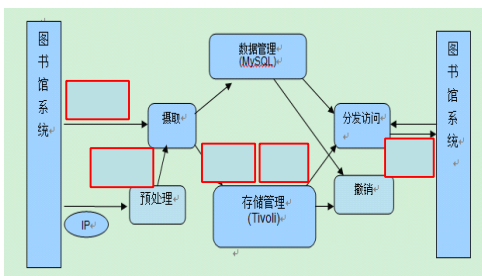


Caplan, Priscilla. DAITSS (Dark Archive in the Sunshine State). Presentation at the Partnerships in Innovation Symposium, National Archives, November 15-16 2004.

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

信息包



第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

存储管理

-混合存储管理

- ✓ 全部元数据存放在 MySQL，转换成 METS 元数据文件与内容文件以 AIP 包形式存储
- ✓ 永久标识符，保存 AIP 与元数据链接

-第三方存储管理系统

- ✓ DAITSS 使用 IBM Tivoli Storage Management
- ✓ DAITSS 调用第三方存储管理功能

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

保存策略

-保存层次

- ✓ 比特层次保存 (bit-level preservation)
- ✓ 完全保存 (full preservation)
- 12种不同文件格式完全保存: AIFF、AVI、JPEG、JP2、JPX、PDF、txt、QuickTime、TIFF、WAVE、XML、XML DTD
- ✓ 保存层次由存档机构指定

-异地多重备份策略

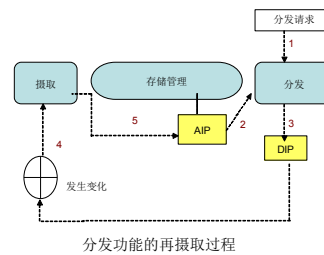
- ✓ AIP 三份备份
- ✓ 两份写入本地磁带设备，一份通过网络在佛罗里达的 Tallahassee 进行异地存储

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

-格式迁移策略

- ✓ 按需迁移策略
- ✓ 摄取模块进行再摄取 (re-ingested) 完成格式迁移



Reference: Chuck Thomas „Preserving ETDs with DAITSS.“

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

– 本地化、规范化保存策略

- ✓ 建立附件文档的本地文件版本
XML的DTD文档
- ✓ 摄入过程对文件进行规范性检验
 - 不存储规范产生的新版本，保证在需要生成规范版本时，能够正确的创建规范文件
 - PDF (DAITSS最佳保存格式)
XML、HTML (DAITSS对内部链接文档规范)

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

应用案例分析



FDA于2005年11月投入使用，资源主要是当地数字化项目中的ETDs（硕博学位论文）。目前FDA提供对12种不同文件格式的完整保存。

FDA2009年2月的摄入统计：

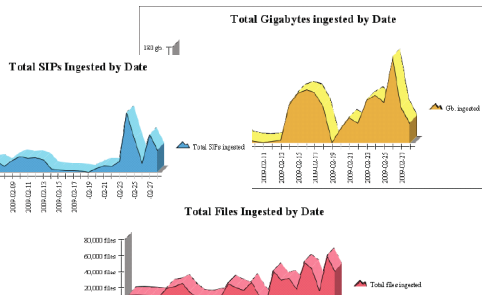
Total Holdings of the FDA as of Mar 1, 2009 12:00 AM:

Packages:	96,127
Files:	9,553,399
Bytes:	19,641,685,005,62

截止2009年3月1日12:00 AM，FDA有96,127个包，9,553,399文件，共19,641,685,005,62字节（约17.9TB）

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14



FDA 2009年2月摄入统计（SIP数量、文件数量、大小）
http://www.fcla.edu/digitalArchive/pdfs/reports/ingest_stats_February_09.pdf

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

思考

- 设计上的优势：专注于保存的后台仓储系统；
- 积极的保存策略：基于规范、迁移、本地化的保存策略；
- DAITSS建立社区，开发人员不断完善和增强系统功能；
- 更方便与其他系统整合，广泛接受和应用性是须考虑的问题。

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14

Further Study

DAITSS software: <http://daitss.fcla.edu/>

Florida Digital Archive:
<http://www.fcla.edu/digitalArchive/index.htm>

吴振新 wuzx@mail.las.ac.cn
向菁 xiangj@mail.las.ac.cn

第二届“数字图书馆与开放源码软件”学术研讨会

2009.3.11—3.14